A Survey of Large Language Models in Medicine: Progress, Application, and Challenge

Abstract

Large language models (LLMs), such as ChatGPT, have received substantial attention due to their capabilities for understanding and generating human language. While there has been a burgeoning trend in research focusing on the employment of LLMs in supporting different medical tasks (e.g., enhancing clinical diagnostics and providing medical education), a review of these efforts, particularly their development, practical applications, and outcomes in medicine, remains scarce. Therefore, this review aims to provide a detailed overview of the development and deployment of LLMs in medicine, including the challenges and opportunities they face. In terms of development, we provide a detailed introduction to the principles of existing medical LLMs, including their basic model structures, number of parameters, and sources and scales of data used for model development. It serves as a guide for practitioners in developing medical LLMs tailored to their specific needs. In terms of deployment, we offer a comparison of the performance of different LLMs across various medical tasks, and further compare them with state-of-the-art lightweight models, aiming to provide an understanding of the advantages and limitations of LLMs in medicine. Overall, in this review, we address the following questions: 1) What are the practices for developing medical LLMs 2) How to measure the medical task performance of LLMs in a medical setting? 3) How have medical LLMs been employed in real-world practice? 4) What challenges arise from the use of medical LLMs? and 5) How to more effectively develop and deploy medical LLMs? By answering these questions, this review aims to provide insights into the opportunities for LLMs in medicine and serve as a practical resource. We also maintain a regularly updated list of practical guides on medical LLMs at https://github.com/AI-in-Health/MedLLMsPracticalGuide